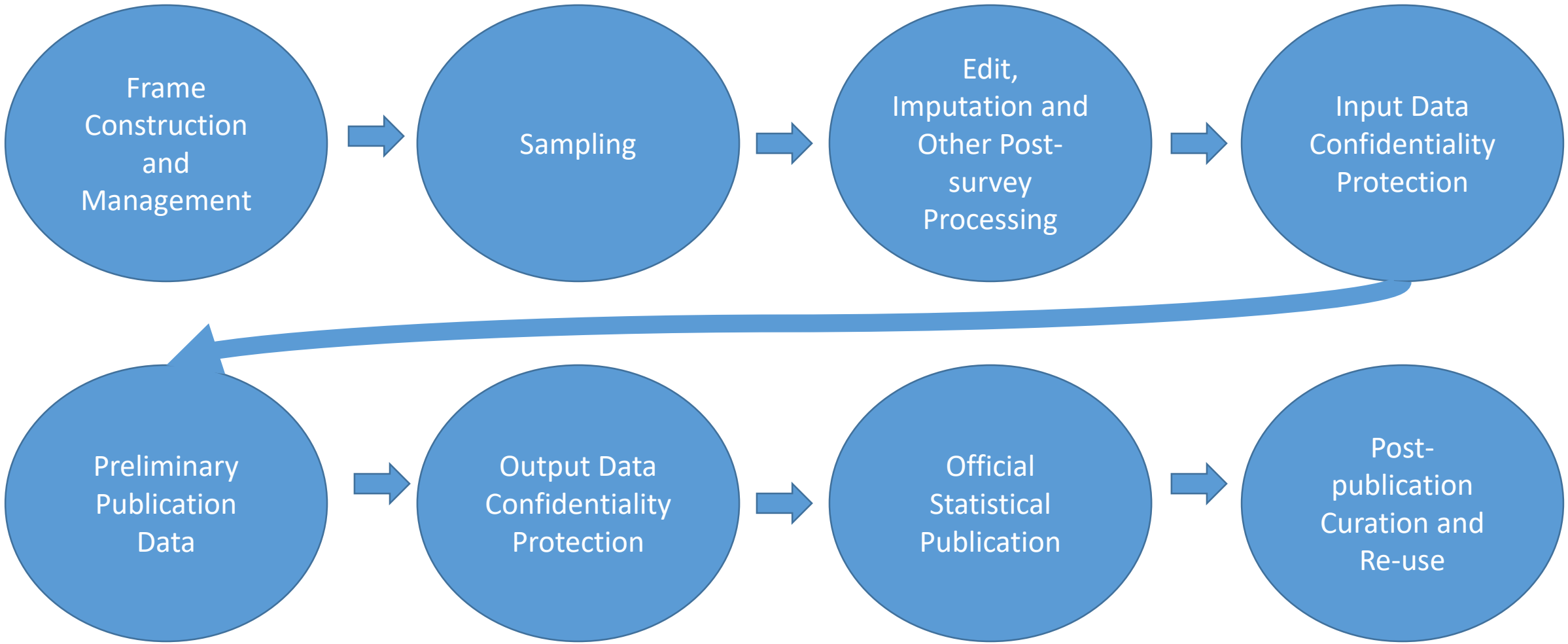
# An Integrated Approach to Statistical Agency Modernization

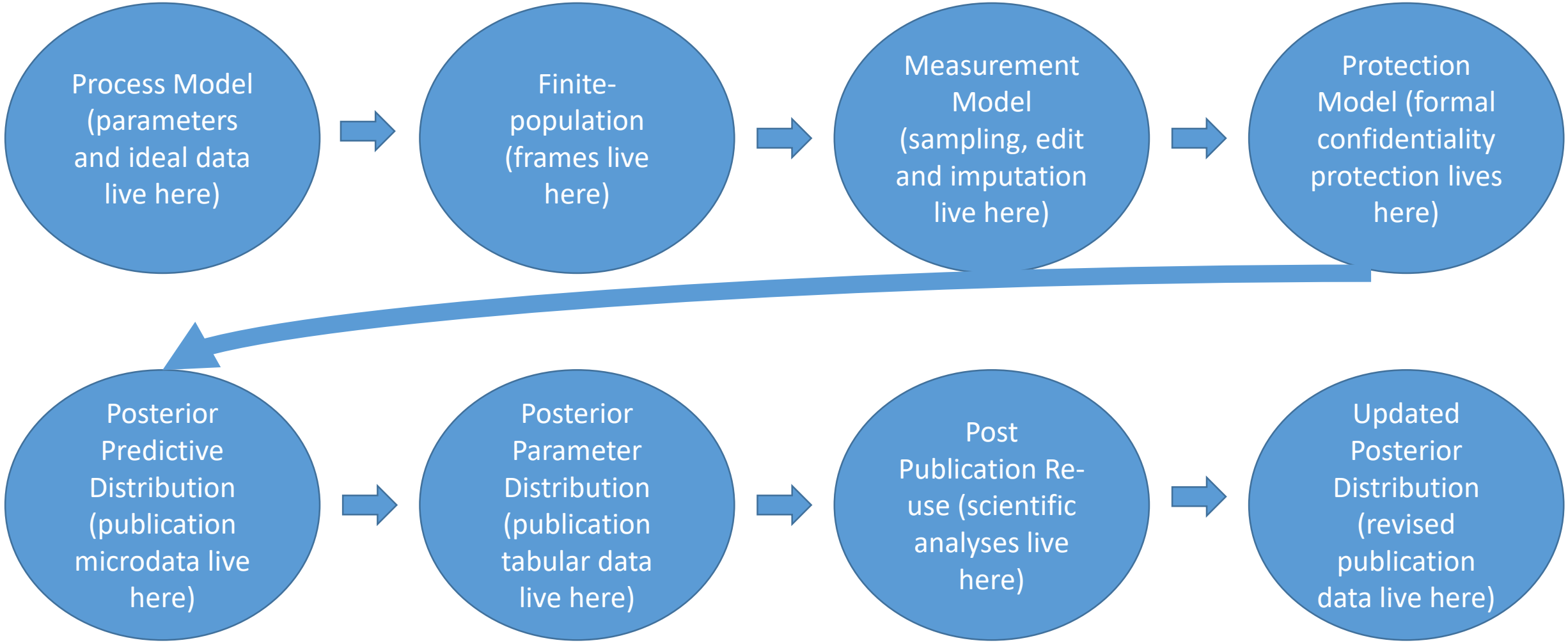
John M. Abowd
Cornell University and U.S Census Bureau*
Workshop on Spatial and Spatio-Temporal Design and Analysis for Official Statistics
May, 2016

*Prepared as a Cornell professor before I assumed an executive role at the Census Bureau. My Census Bureau appointment begins on June 1, 2016.

# Current Data Manufacturing Process


Frame Construction and Management
Sampling
Edit, Imputation and Other Post-survey Processing
Input Data Confidentiality Protection
Preliminary Publication Data
Output Data Confidentiality Protection
Official Statistical Publication
Post-publication Curation and Re-use

# Hierarchical Approach to Data Manufacturing


Process Model (parameters and ideal data live here)
Finite-population (frames live here)
Measurement Model (sampling, edit and imputation live here)
Protection Model (formal confidentiality protection lives here)
Posterior Predictive Distribution (publication microdata live here)
Posterior Parameter Distribution (publication tabular data live here)
Post Publication Re-use (scientific analyses live here)
Updated Posterior Distribution (revised publication data live here)

# Example 1: Monthly Retail Trade

# Monthly & Annual Retail Trade

## Advance Monthly Retail Trade Report

The **April 2016** Advance Monthly Sales for Retail Trade and Food Services report was released on May 13, 2016 at 8:30 a.m., and available as:

- Full Publication in Excel [249KB] | PDF [384KB]
- Time Series (Adjusted Sales Data/Seasonal Factors—1992 to present)

**Time Series/Trend Charts:** Create your own customizable time series. NEW

## Monthly Retail Trade Report

The **March 2016** Monthly Retail Trade and Food Services report was released on May 13, 2016 at 8:30 a.m. for sales and 10:00 a.m. for inventories, and available as:

- Retail and Food Services Sales: Excel (1992-present) [633KB]
- Retail Inventories and Inventories/Sales Ratios: Excel (1992-present) [404KB]
- Adjustment Factors for Seasonal and Other Variations of Monthly Estimates: Sales | Inventories
- Reliability of Monthly Estimates: Sales | Inventories
- Annual Revision of Monthly Retail and Food Services: Sales and Inventories--January 1992 through March 2016

**Time Series/Trend Charts:** Create your own customizable time series. NEW

## Latest Annual Retail Trade Report

The 2014 Annual Retail Trade Report was released on March 7, 2016. A Summary of Changes provides comparability with previous surveys.

- **Annual Retail Trade Survey—2014**:
  Sales (1992-2014): Excel [66KB]
  Sales Taxes (2004-2014): Excel [46KB]
  Inventories (1992-2014): Excel [44KB]
  Purchases (1992-2014): Excel [46KB]

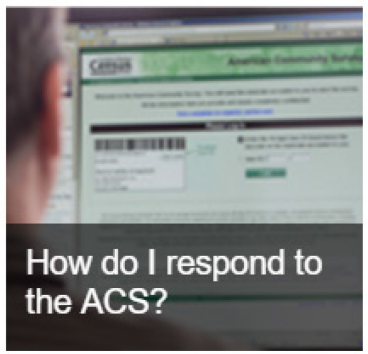
Monthly & Annual Retail Trade, Main Page - US Census Bureau - Google Chrome

# Example 2: American Community Survey

# American Community Survey (ACS)



The American Community Survey helps local officials, community leaders and businesses understand the changes taking place in their communities. It is the premier source for detailed information about the American people and workforce.

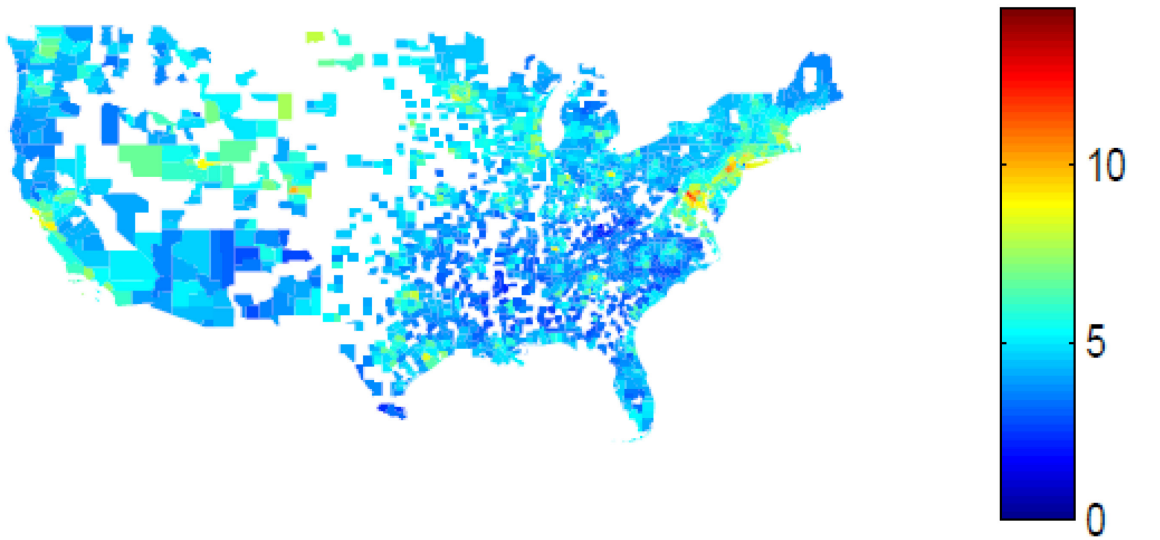





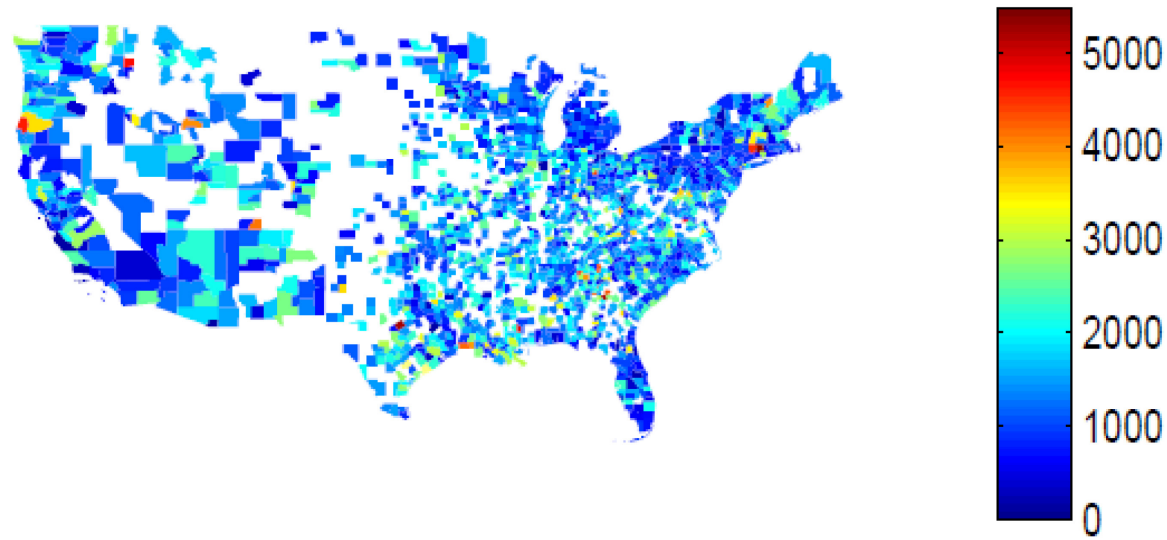
## Latest

Data | News | Events | Library

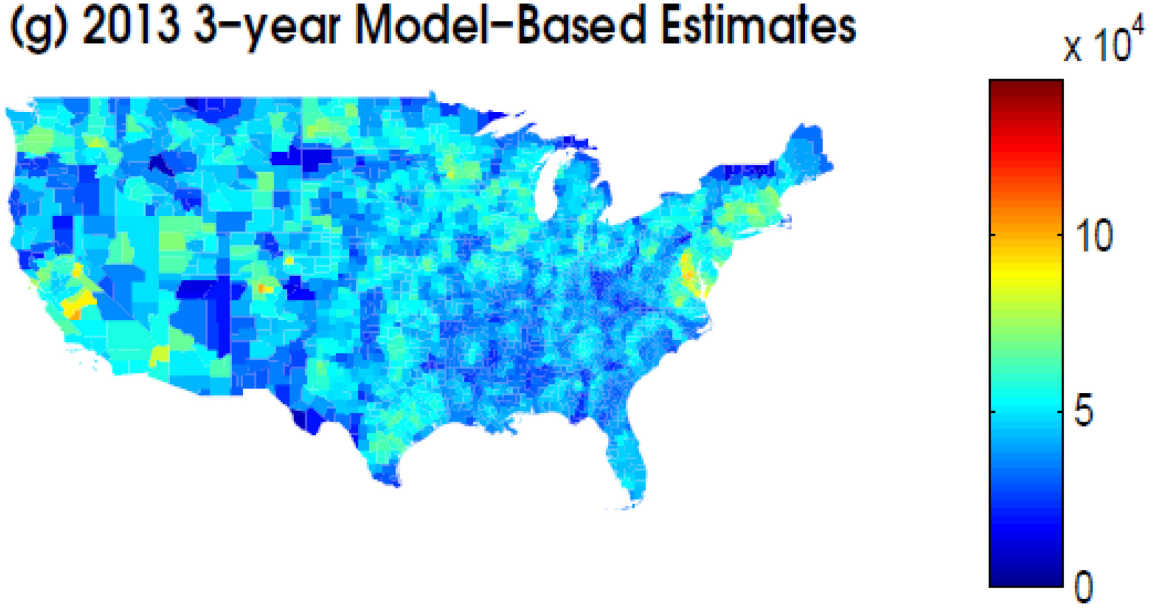
(c) 2013 3-year ACS Estimates



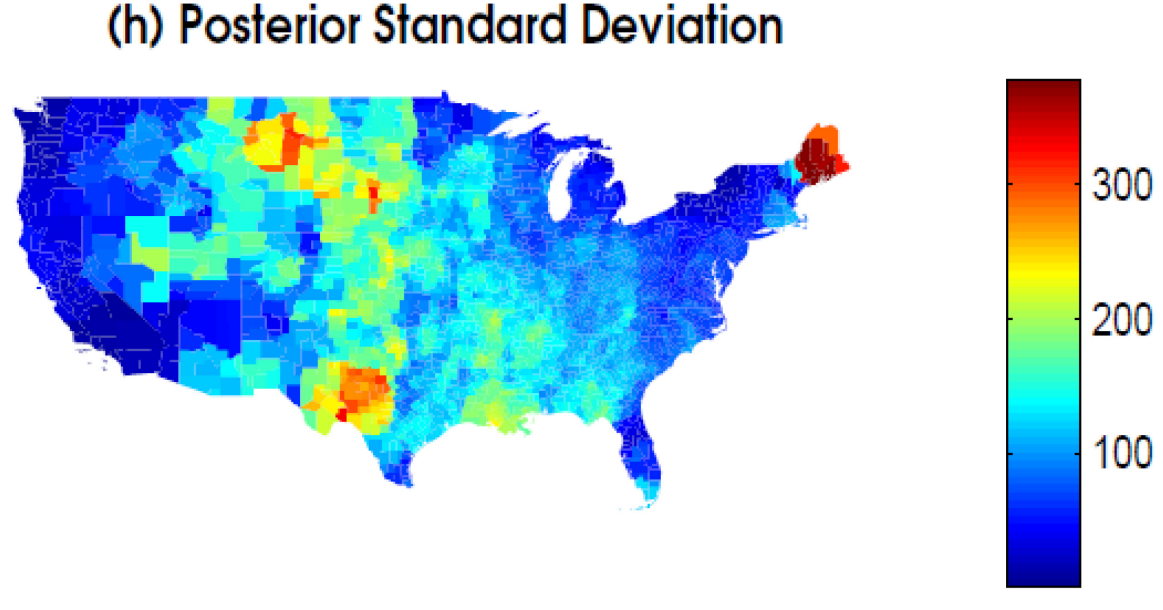

(d) 2013 3-year ACS Estimates of Std.Dev




(g) 2013 3-year Model-Based Estimates




(h) Posterior Standard Deviation



## MANHATTAN

**1,2 WALL STREET, CIVIC CENTER, GOVERNORS ISLAND, LIBERTY ISLAND, ELLIS ISLAND, TRIBECA, GREENWICH VILLAGE, NOHO, SOHO, LITTLE ITALY**

- 42 percent more households with children
- 54 percent more people age 55 to 64
- 137 percent more residents who work in protective services (police, security, etc.)

**3 LOWER EAST SIDE, CHINATOWN**

- 55 percent more adults with bachelor's degrees but no higher degrees
- 43 percent more households of men living alone
- 24 percent fewer Hispanic residents

**4,5 CHELSEA, HELL'S KITCHEN, HERALD SQUARE, MIDTOWN, TIMES SQUARE**

- 30 percent more residents age 35 to 44
- 21 percent fewer households headed by women
- 52 percent fewer homes owned and occupied by Hispanics

**6 MURRAY HILL, EAST MIDTOWN, STUYVESANT TOWN**

- 33 percent fewer adults with some college education but no four-year degrees
- 42 percent fewer residents who work in transportation
- 15 percent fewer residents who are widowed, divorced or separated

**7 UPPER WEST SIDE, LINCOLN SQUARE**

- 34 percent fewer Hispanic families
- 24 percent more married residents
- 105 percent more children under 5


THE NEW YORK TIMES

- 46 percent fewer residents who work in construction and manufacturing

**2 SUNNYSIDE, WOODSIDE**

- 29 percent more residents age 55 to 64
- 39 percent fewer residents who work in construction and manufacturing
- 17 percent more residents who are widowed, divorced or separated

**3 JACKSON HEIGHTS, EAST ELMHURST, NORTH CORONA**

- 29 percent fewer black households

**4 ELMHURST, CORONA**

- 36 percent fewer blacks
- 24 percent fewer households of women living alone

**5 MASPETH, RIDGEWOOD, MIDDLE VILLAGE, GLENDALE**

- 56 percent more residents who work in health care
- 14 percent fewer households of women living alone
- 26 percent fewer residents with less than a high school education

**6 REGO PARK, FOREST HILLS**

- 47 percent more residents who work in building services (janitors, superintendents, etc.)
- 41 percent more residents who work in restaurants and food services
- 48 percent fewer black families

**7 FLUSHING, WHITESTONE, COLLEGE POINT**

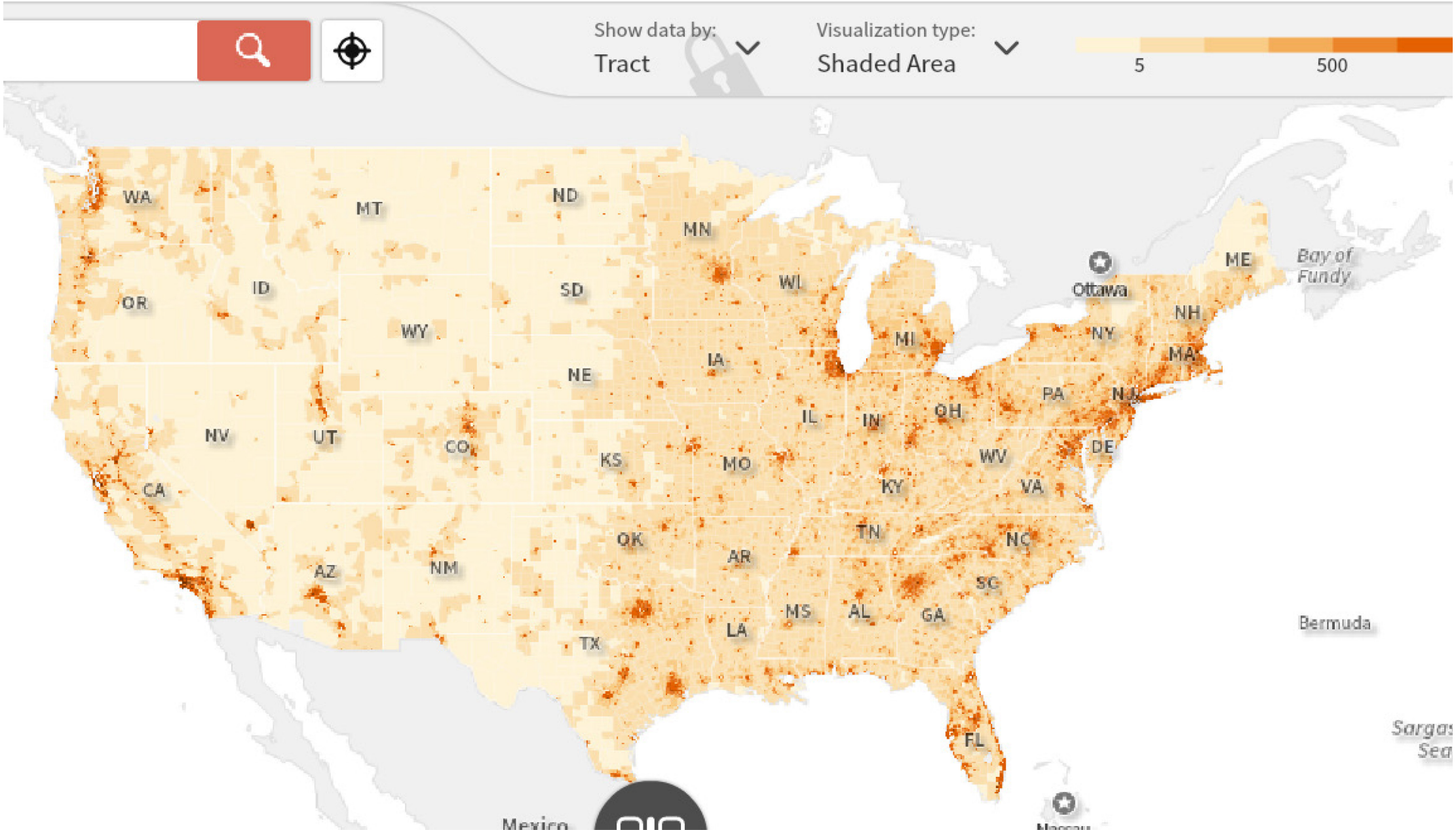
- 45 percent more employed workers
- 50 percent more residents who work in personal services

**8 FRESH MEADOWS, JAMAICA HILLS, KEW GARDENS HILLS**

- 24 percent more households of men living alone

# Population Density (per sq. mile)

ACS 2014 (5-Year Estimates)



# Example 3: RAPPOR

# RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response

Úlfar Erlingsson
Google, Inc.
ulfar@google.com

Vasyl Pihur
Google, Inc.
vpihur@google.com

Aleksandra Korolova
University of Southern California
korolova@usc.edu

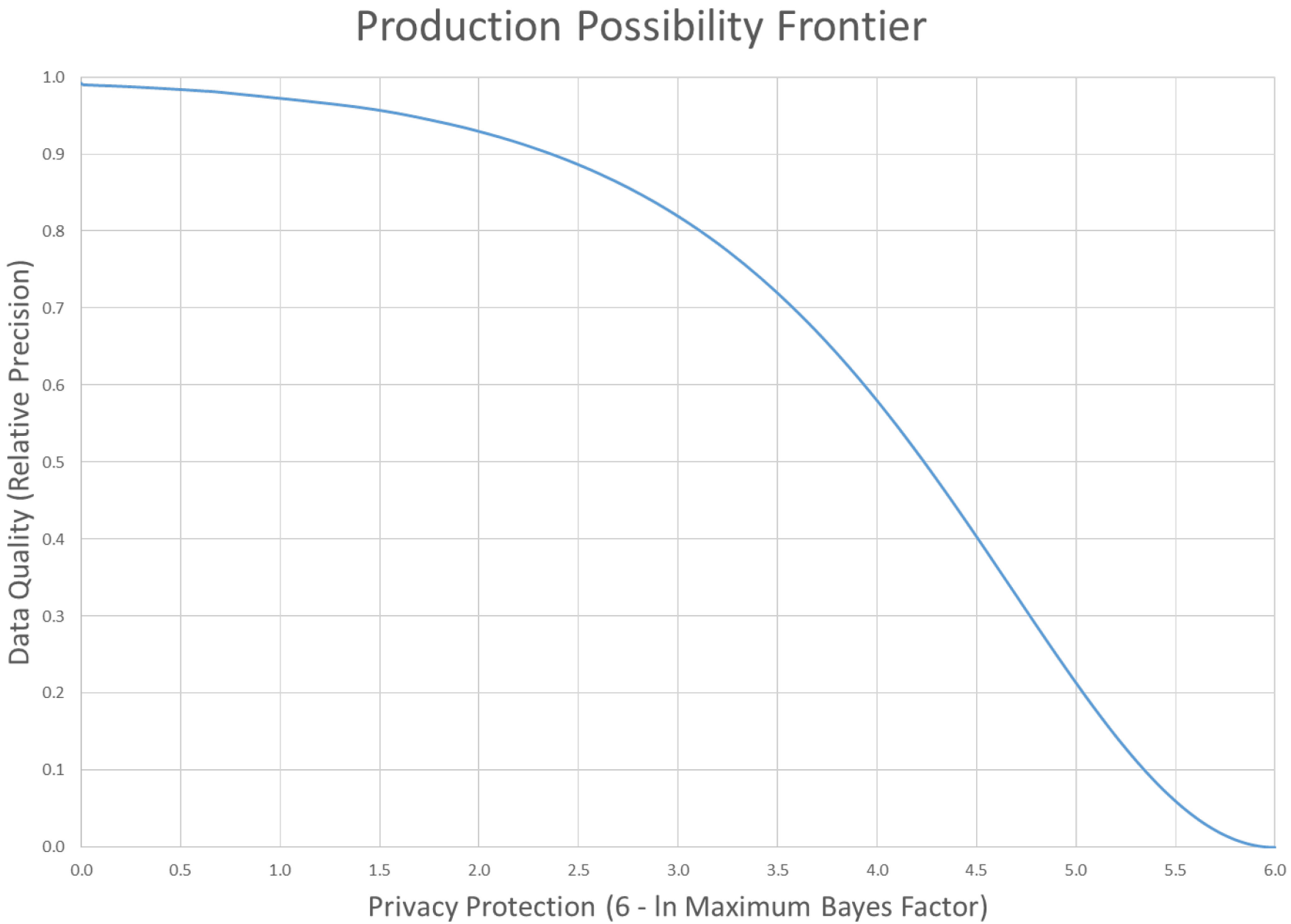
## ABSTRACT

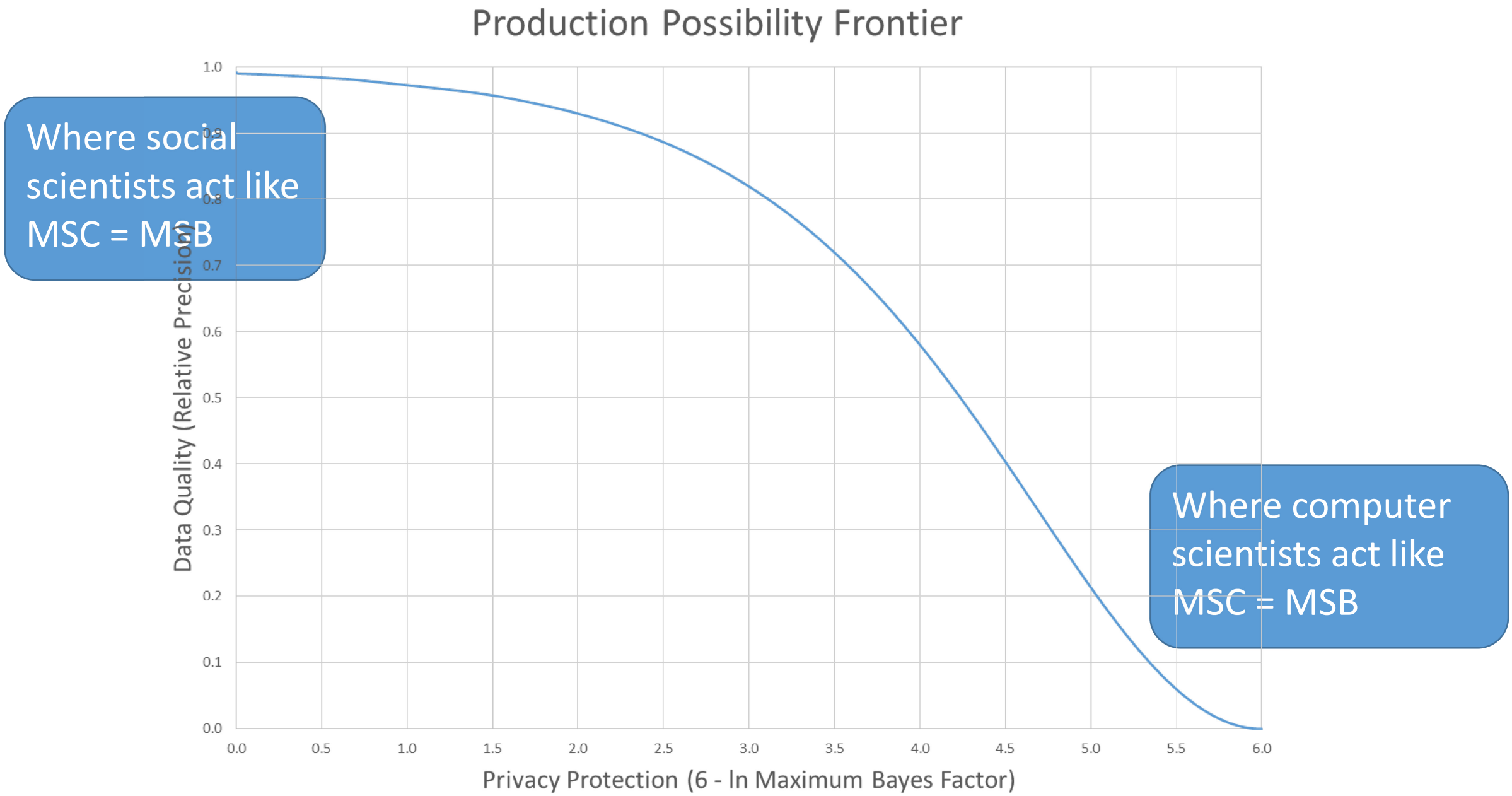
Randomized Aggregatable Privacy-Preserving Ordinal Response, or RAPPOR, is a technology for crowdsourcing statistics from end-user client software, anonymously, with strong privacy guarantees. In short, RAPPORs allow the forest of client data to be studied, without permitting the possibility of looking at individual trees. By applying randomized response in a novel manner, RAPPOR provides the mechanisms for such collection as well as for efficient, high-utility analysis of the collected data. In particular, RAPPOR permits statistics to be collected on the population of client-side strings with strong privacy guarantees for each client, and without linkability of their reports.

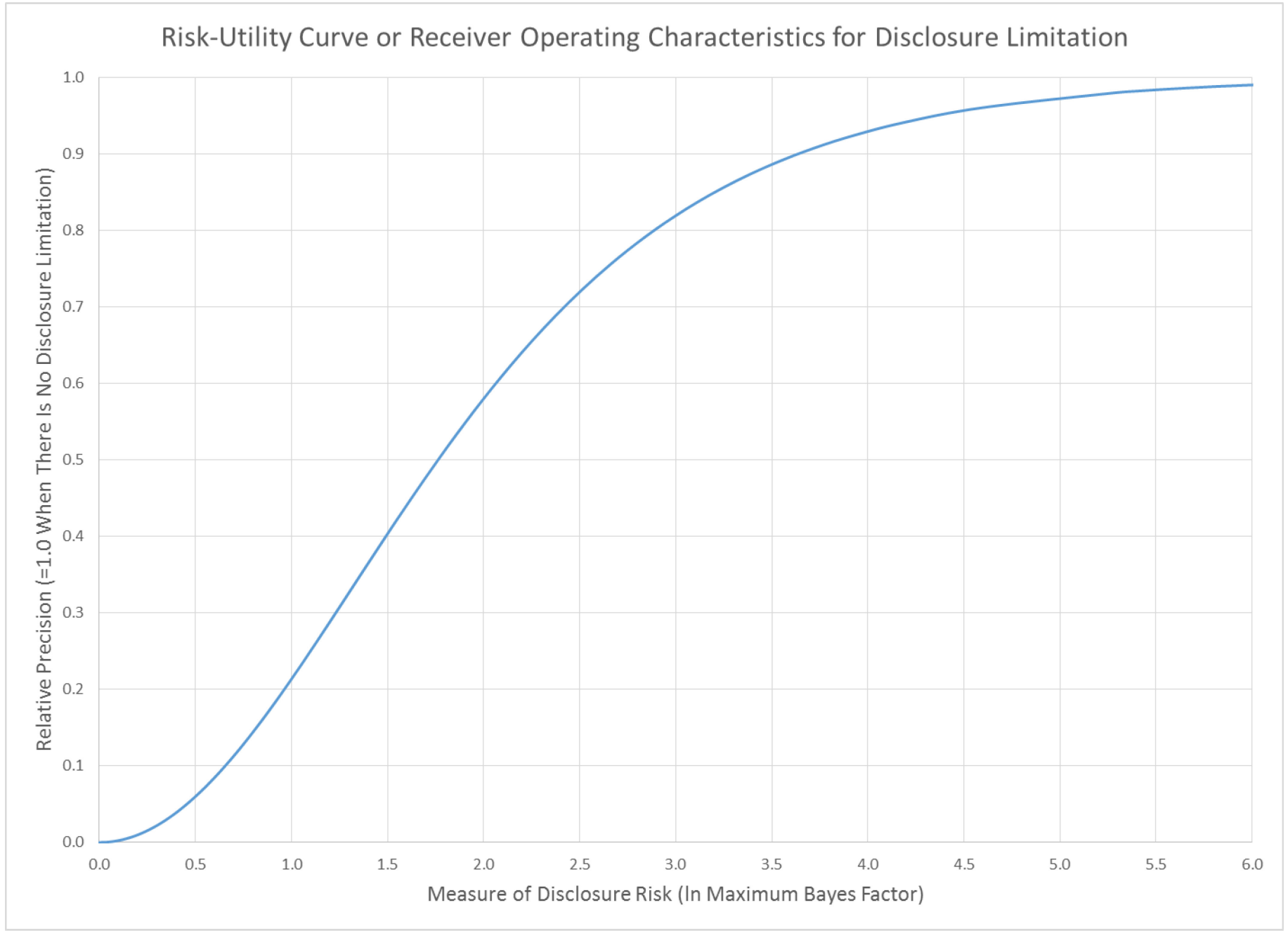
This paper describes and motivates RAPPOR, details its differential-privacy and utility guarantees, discusses its practical deployment and properties in the face of different attack models, and, finally, gives results of its application to both

asked to flip a fair coin, in secret, and answer "Yes" if it comes up heads, but tell the truth otherwise (if the coin comes up tails). Using this procedure, each respondent retains very strong deniability for any "Yes" answers, since such answers are most likely attributable to the coin coming up heads; as a refinement, respondents can also choose the untruthful answer by flipping another coin in secret, and get strong deniability for both "Yes" and "No" answers.

Surveys relying on randomized response enable easy computations of accurate population statistics while preserving the privacy of the individuals. Assuming absolute compliance with the randomization protocol (an assumption that may not hold for human subjects, and can even be non-trivial for algorithmic implementations [23]), it is easy to see that in a case where both "Yes" and "No" answers can be denied (flipping two fair coins), the true number of "Yes" answers can be accurately estimated by $2(Y - 0.25)$, where


Production Possibility Frontier
Data Quality (Relative Precision)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Privacy Protection (6 - ln Maximum Bayes Factor)


Production Possibility Frontier
Where social scientists act like MSC = MSB
Where computer scientists act like MSC = MSB
Data Quality (Relative Precision)
Privacy Protection (6 - ln Maximum Bayes Factor)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0


Risk-Utility Curve or Receiver Operating Characteristics for Disclosure Limitation
Relative Precision (=1.0 When There Is No Disclosure Limitation)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Measure of Disclosure Risk (In Maximum Bayes Factor)

# Example 4: SIPP Synthetic Data

# Survey of Income and Program Participation

About this Survey

Information for Respondents

Data

Events

**Guidance for Data Users**

DataFerrett

SIPP FTP site

**Synthetic SIPP Data**

SIPP Users' Guide

Methodology

News

Publications

Technical Documentation

Working Papers

## Synthetic SIPP Data



### Background on the SIPP Synthetic Beta

The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receip and were produced by Census Bureau staff economists and statisticians in collaboration with researchers at Cornell University, the SSA and the IRS. The purpose of the SSB is to provide access to linked data that are usually not publically available due to confidentiality concerns. To overcome these concerns, Census synthesizes, or models, all the variables in a way that changes the record of each individual so as to preserve the underlying covariate relationships between the variables. Only gender and a link to the first reported marital partner are not altered by the synthesis process and still contain their original values.

Nine SIPP panels (1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, and 2008) form the basis for the SSB, with a subset of variables available across all the panels selected for inclusion and harmonization of variable definitions across the years covered by the panels. Administrative data are added and some editing is done to correct for logical inconsistencies in the IRS and Social Security earnings and benefits data. Thus, the SSB is a particularly appealing data set for new SIPP users because little data preparation is needed. A complete list of variables included in SSB version 6.0, along with details about the harmonization and editing, is available in our Codebook.

As part of the synthesis process, missing survey data and missing administrative data were multiply imputed. The resulting data sets are called the Completed Gold Standard Files and contain all original, non-missing, confidential values and imputed values in place of originally missing data. These files form the basis for evaluating results from the synthetic data. The goal of the SSB is to produce results that are qualitatively the same as results from the Completed Gold Standard Files.

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

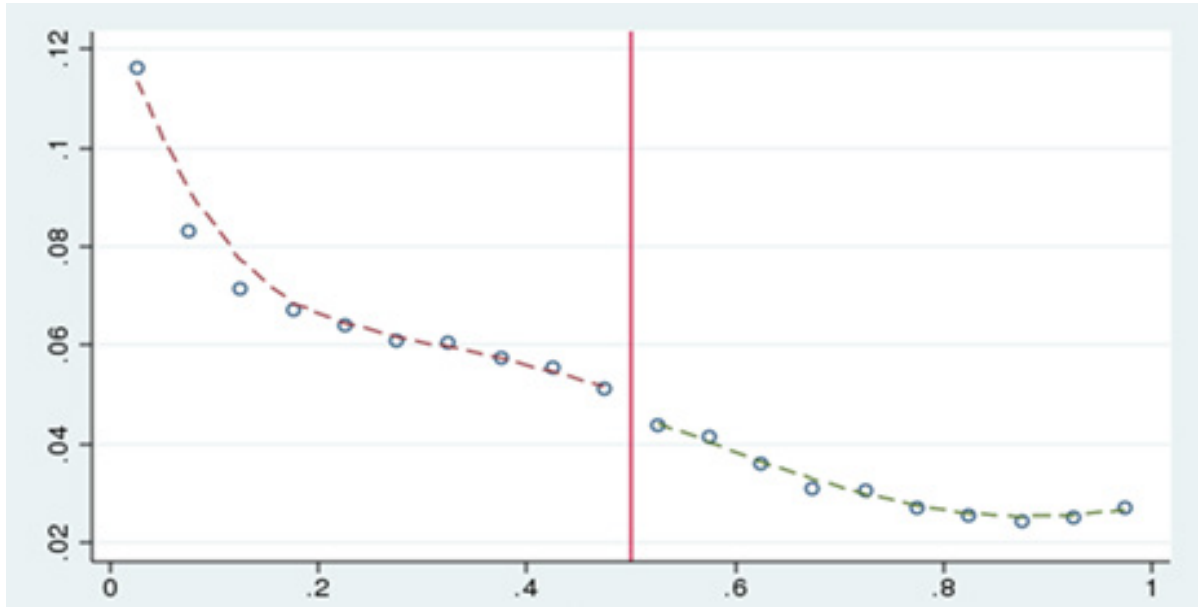
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

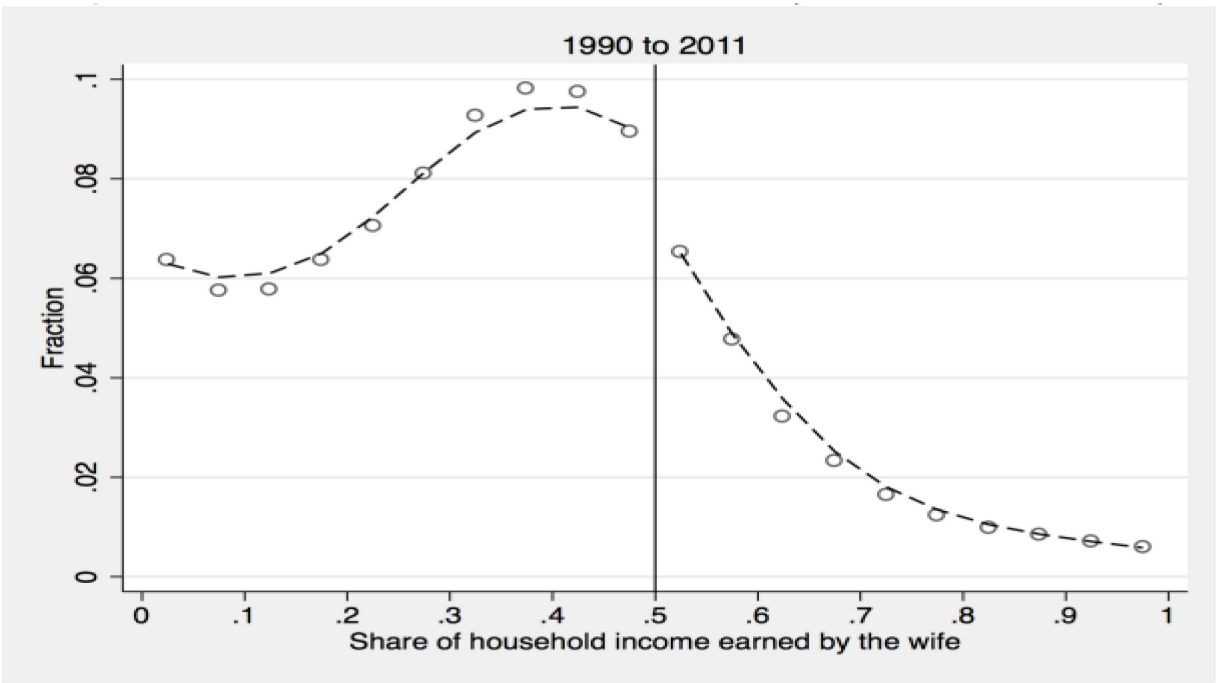
# AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

*Provides Principles to Improve the Conduct and Interpretation of Quantitative Science*

March 7, 2016

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*




Bertrand, Kamenica and Pan (QJE 2015), doi: 10.1093/qje/qjv001

# Thank you.

Contact: john.maron.abowd@census.gov